Split-Brain Harness

AI Security Infrastructure for Government LLM Deployments

SGAIL · North Shore, Oahu, HI · DHS SBIR Phase 1 Candidate

```
sbh demo --serve --offline   # live demo — no backend required
```

Rust · MIT · Single static binary · Air-gap ready

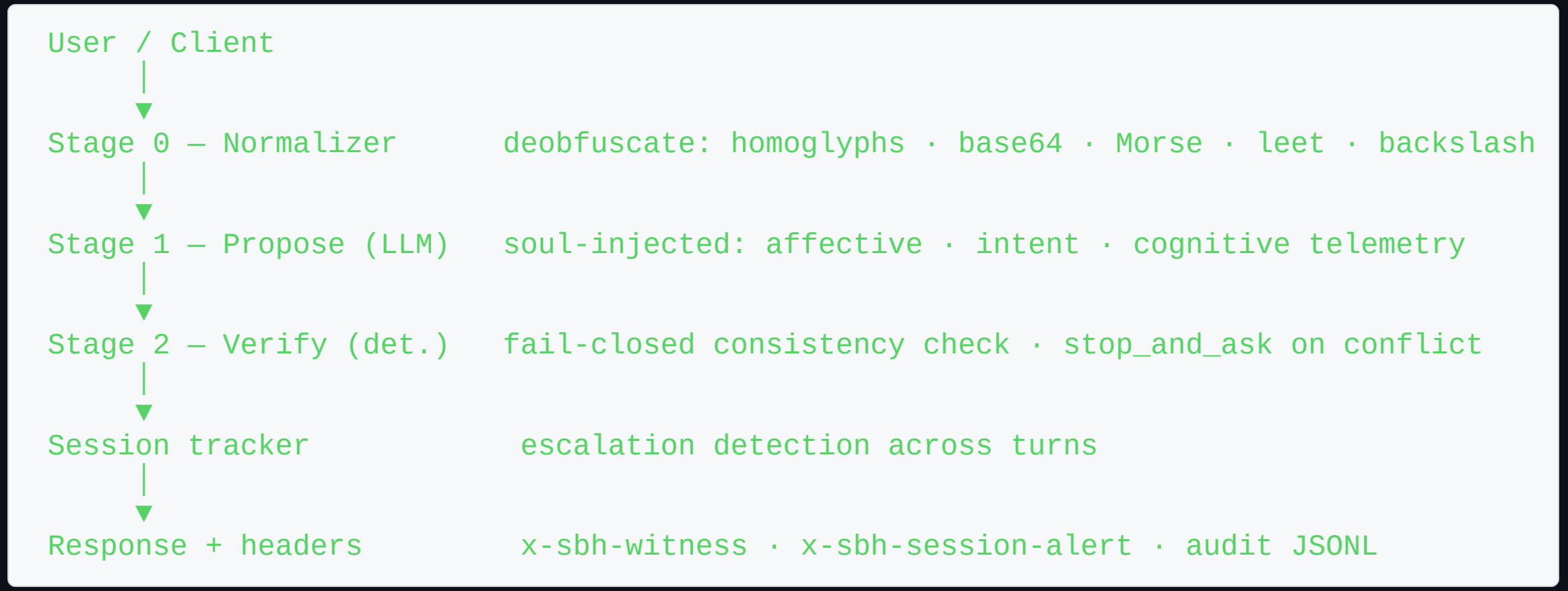
The Problem

LLMs are being deployed in cleared and government-adjacent environments with **no security telemetry**.

Attack vector	What the model sees	What it actually is
Prompt injection	Legitimate user request	Instruction override targeting keys/config
Insider threat	"Help me map data access"	Audit-gap reconnaissance
Authority impersonation	"Director Hargrove, NSA..."	Fabricated federal authority
Encoding evasion	`ignore all instructions`	Cyrillic homoglyphs → payload
Slow-boil escalation	3 benign turns, then attack	Missed without session context

Split-Brain Harness

Drop-in OpenAI-compatible proxy. Every request passes through a **two-stage telemetry pipeline** before reaching the model.



Benchmark Results

Evaluated on three public adversarial datasets · **llama3.2:3b · local Ollama · air-gapped**

Dataset	Rows	Precision	Recall	F1	Notes
CyberEC	141	**1.00**	0.50	0.67	Zero false positives
TrustAI Jailbreaks	1,398	—	—	**94.8% flagged**	Unlabeled
Deepset Prompt Injections	546	0.81	0.37	0.51	3B local model limit†

Stage 0 normalizer catches 50% of CyberEC encoding-evasion FNs: homoglyphs · base64 · Morse · backslash-escape · fullwidth · leet

† Deepset recall improves with capable backend (Claude/GPT-4).

Key Capabilities

Capability	SBH	Standard LLM Gateway
Two-stage telemetry (propose + verify)	✅	—
Stage 0 deobfuscation normalizer	✅	—
Soul-injected identity baseline	✅	—
Multi-turn session escalation detection	✅	—
Air-gap / local model capable	✅	rarely
Ephemeral sandboxed tool execution (WASM)	✅	—
OpenAI-compatible drop-in proxy	✅	varies
JSONL audit trail + Prometheus metrics	✅	varies
Single static Rust binary	✅	—

DHS SBIR Phase 1

What we're building

A hardened, air-gap-deployable AI security layer for government LLM deployments — tamper-evident audit trail, session-level threat detection, sandboxed tool execution.

Phase 1 milestones (~$300K · 6 months)

Milestone	Deliverable
M1	Benchmark suite against DHS-relevant labeled datasets
M2	Hardened `sbh serve` with FedRAMP-aligned audit controls
M3	Normalizer v2 — full Unicode TR39 confusables, entropy scoring
M4	Red-team evaluation by independent cleared assessor
M5	Open-source release + technical report